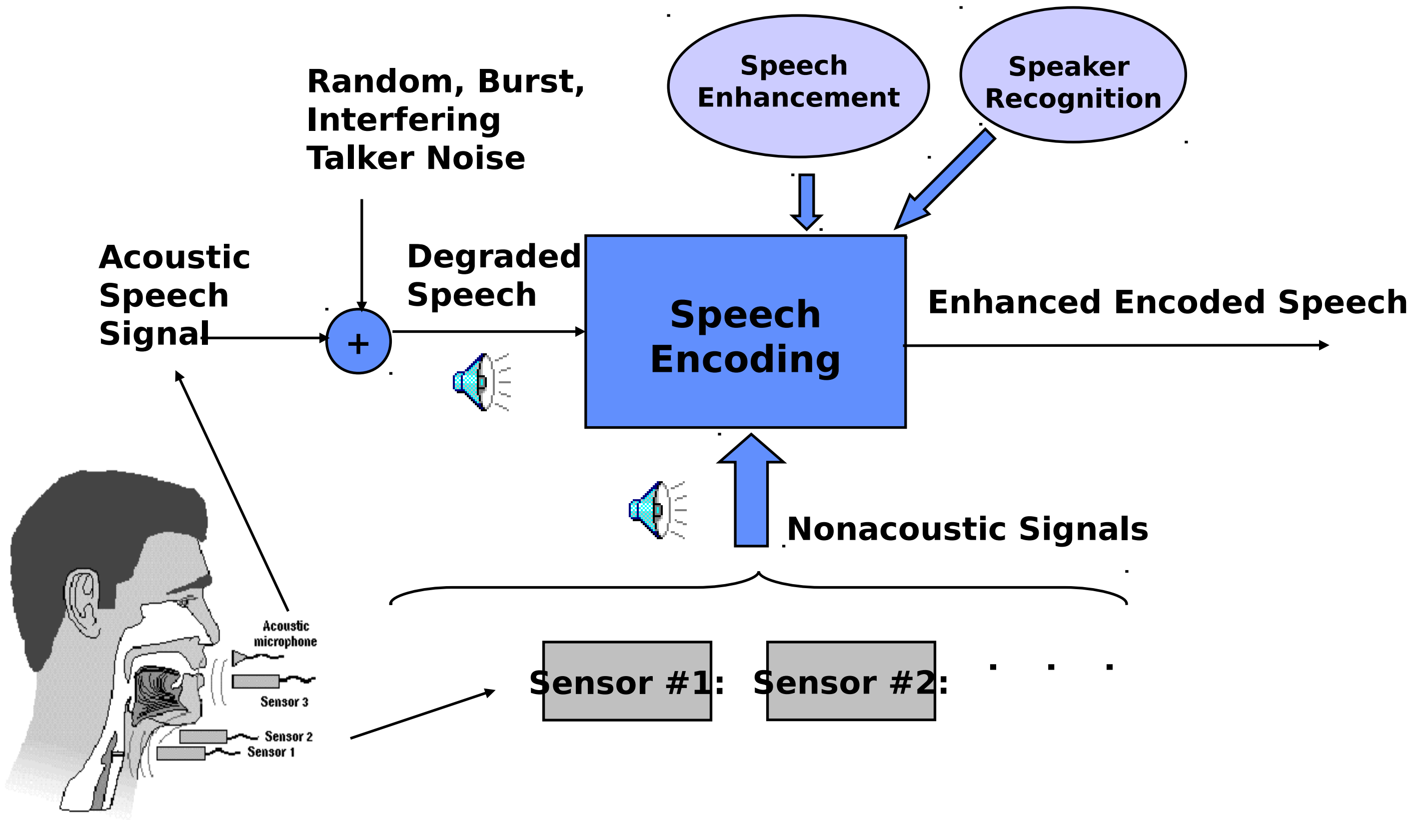
Multisensor Speech Processing

General ASE Objectives

Objective: Use nonacoustic sensors to improve performance of speech encoding algorithms

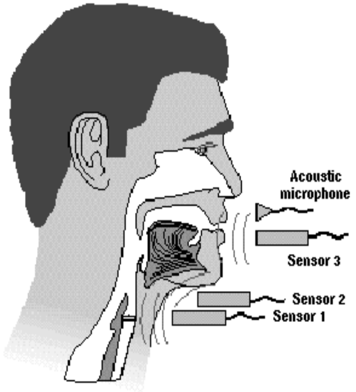
with speech that is degraded by severe additive noise backgrounds

Two Phases: I: 2400 bps and II: 1000-300 bps



Exploiting Nonacoustic Sensors
Primary Phase I Contributions

- **Provided technical support to DARPA and ASE community**
  - **Collaborated with ARCON on Pilot Corpus**
  - **Provided our studies and results to ASE community**
  - **Designed and maintained ASE website**
- **Research and concept studies**
  - **Focus on fundamental issues and theoretical bounds**
  - **Nature of sensor measurements**
    - **Example: Voice bars**
  - **Low-rate speech encoding at 2400 bps**
    - **Fusing acoustic and nonacoustic signals gives large gains in speech intelligibility**
  - **Speaker recognition**
    - **Fusing outputs from different recognition systems using acoustic- and nonacoustic-sensor signals significantly improves speaker recognition accuracy**



Outline

- Nature of sensor measurements in noise
- Applications
  - Speech coding
  - Speaker recognition
- Summary

ASE Corpora
Collection Scenario

- **ASE Corpus Collection performed by ARCON and designed by ARCON and MIT LL**

- **Support R&D and evaluations relevant to ASE**
  - **10 male + 10 female talkers**
  - **Vowel, word, sentence, and conversational material**
    - **Supports DRT and DAM testing**
  - **Primary harsh acoustic noise conditions**
    - **Bradley Tank (M2)**
    - **Black Hawk Helicoptor (BH)**
    - **Military Urban Terrain (MOUT)**

    High and low noise

- **Multisensor recordings of simultaneous channels**
  - **Acoustic microphone used in field: mouth**
    - **Referred to as the *resident microphone* and is a gradient noise-canceling microphone**
  - **B&K reference microphone: mouth**
  - **General Electromagnetic Movement Sensor (GEMS): throat**
  - **1 Electroglottograph (EGG) sensor: throat**
  - **2 P-Mics: throat and forehead**
  - **1 Bone conduction microphone: top of skull (only in MOUT)**

Nonacoustic Sensors

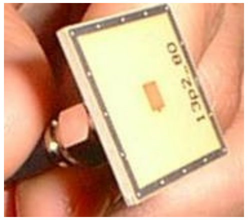
Time-Frequency Properties in Noise

- **For each sensor (GEMS, P-mic and Bone-mic), investigated properties of its signal output in the time and frequency domains**
  - **Source components**
    - **Voicing**
    - **Frication**
    - **Consonant bursts**
  - **Vocal tract components**
    - **Formant location and bandwidth**

- **Time-frequency properties were studied in relation to those of corresponding acoustic microphone signal and to each other in harsh acoustic backgrounds**
  - **Complementary nature of the measurements**

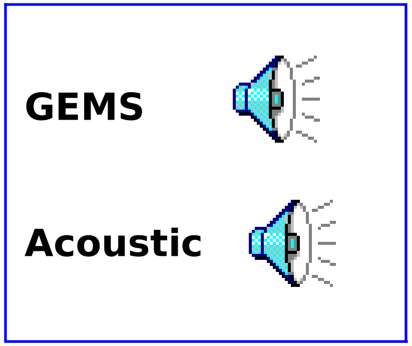
GEMS Nonacoustic Sensor

Signal Properties in Noise



- **In ASE corpus, GEMS placed at the vocal cord location**
- **Observed GEMS signal properties**
  - **Good low-frequency source content**
    - **Low-frequency voicing, including voice bars and nasality**
  - **Strong "glottalized" source activity in low-energy regions**
    - **Irregular pulses at end of words**
    - **Secondary pulses between primary**
  - **Essentially no vocal tract content**
  - **Excellent noise immunity**

Bradley Tank Environment

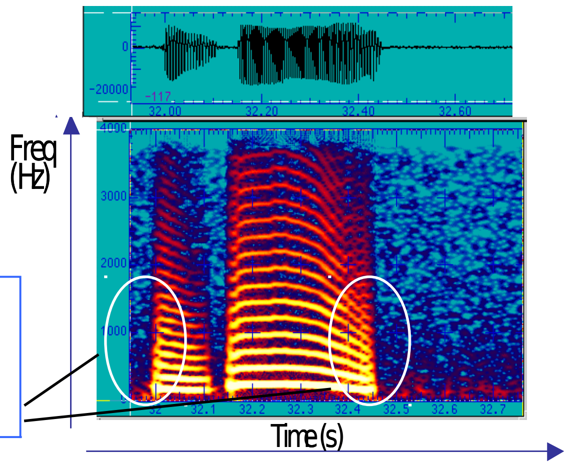


GEMS Nonacoustic Sensor

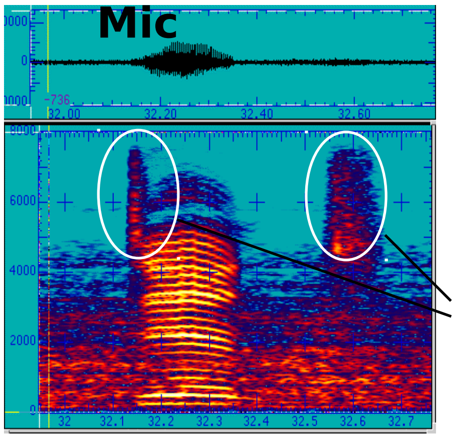
Low-Frequency Voicing

- **Example of low-frequency voicing**
  - **Waveforms (from the Bradley environment) and spectrograms of the resident-mic signal and GEMS signal for the word "<u>d</u>int"**

GEMS



Resident-Mic



The GEMS signal shows the presence of the nasal /n/ and voice bar* in the initial voiced plosive /d/

While the resident-mic shows the high-frequency burst energy in the /d/ and in the unvoiced plosive /t/

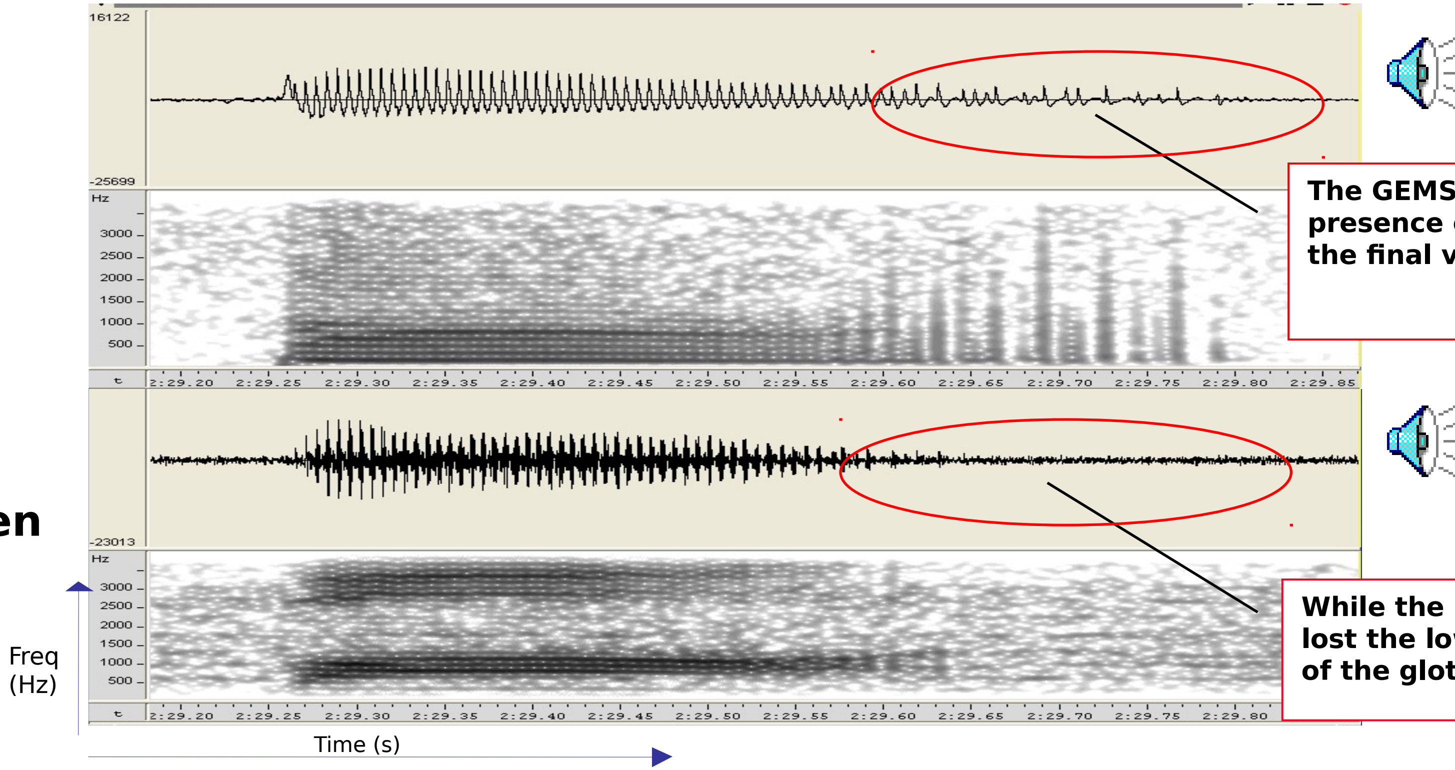
***Voice bars and voicing at the end of consonants is measurable by the GEMS, P-Mic, and Bone-mic. The strength and duration of these vibrations appears to be speaker-dependent, as well as condition-dependent, being more present with harsher noise.**

GEMS Nonacoustic Sensor
Glottalization

- **Example of glottalization**
  - **Waveforms (from the Bradley tank environment) and spectrograms of the resident-mic signal and GEMS signal**

GEMS

Resident-Mic



The GEMS signal shows the presence of glottalization in the final vowel

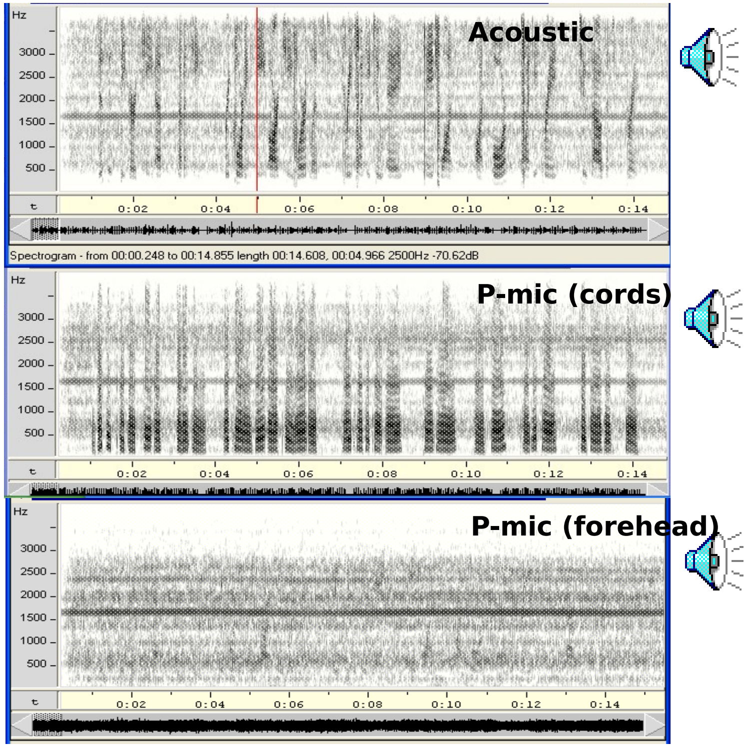
While the resident-mic has lost the low-energy portion of the glottalized signal

P-Mic Nonacoustic Sensor

Signal Properties in Noise

- **In ASE corpus, P-Mic placed near vocal cords and forehead**
- **Observed P-Mic signal properties near vocal cords**
  - **Good low-frequency source content**
    - **Low-frequency voicing, including voice bars and nasality**
  - **Some glottalized source activity in low-energy regions**
  - **Some vocal tract content**
  - **Fair noise immunity**
- **Observed P-Mic signal properties on forehead**
  - **Good source and tract content**
  - **Poor noise immunity**

- **Example**
  - **Black Hawk Helicopter Environ**

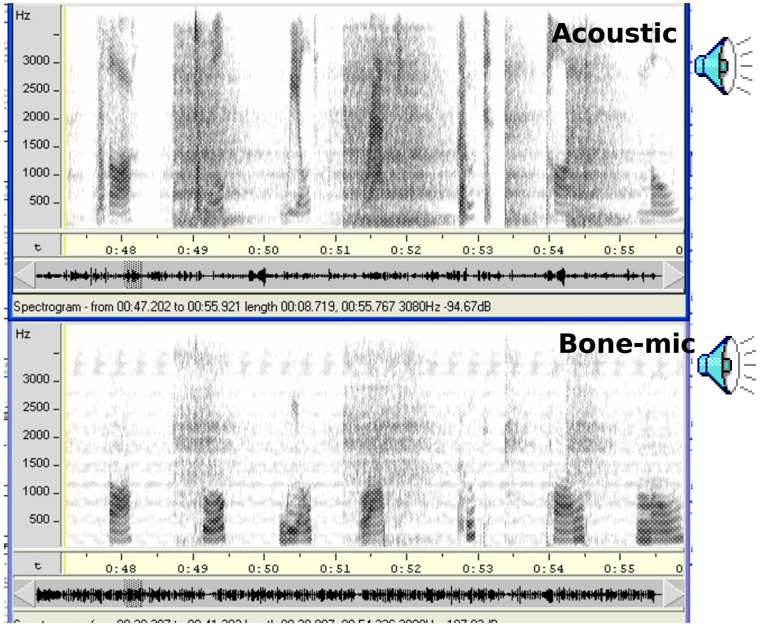


Bone-Conduction Nonacoustic Sensor

Signal Properties in Noise

- **In ASE corpus, bone-conduction mic placed on skull**
- **Observed bone-mic signal properties**
  - Good mid-frequency spectral content
  - Fair glottalized source activity in low-energy regions
  - Good vocal tract content
  - Good noise immunity

- **Example**
  - Black Hawk Helicopter Environ



Fundamental Measurements

Approximate Sensor Contributions

Sensor / Contribution	Resident Microphone	GEMS (Glottally-located)	P-mic* (Throat-located)	Bone-mic (Head-located)
Voice Bars				
Voiced Speech				
Unvoiced Speech				
Low-frequency Vocal Tract				
High-frequency Vocal Tract				

High Quality | Low Quality | Poor Quality

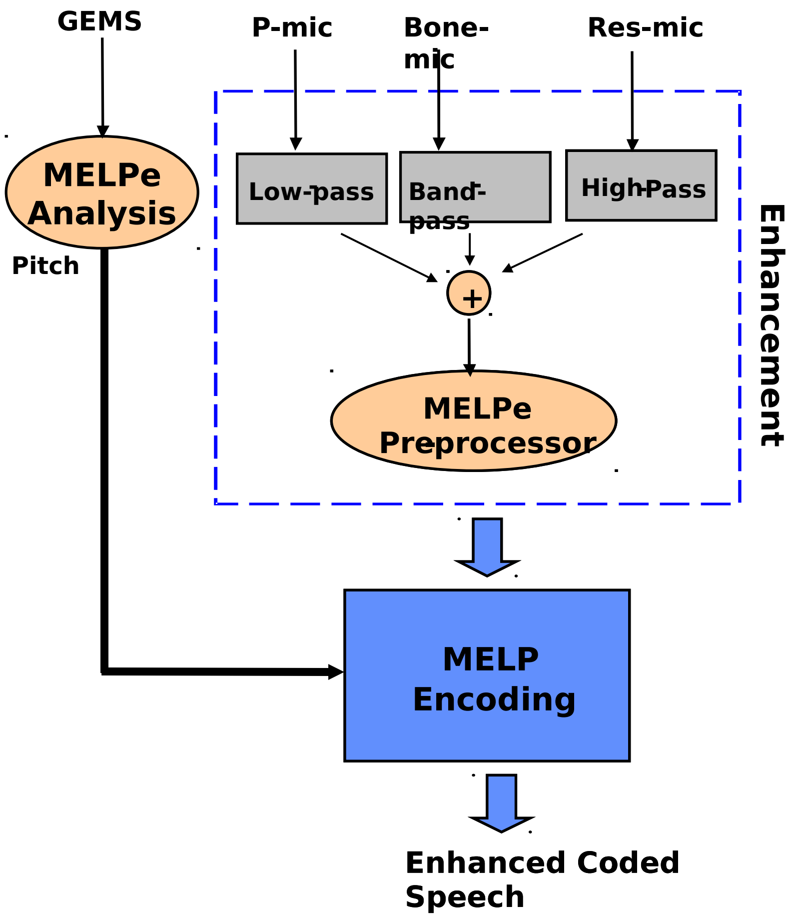
*Second P-mic at forehead

Outline

- Nature of sensor measurements in noise
- Applications
  - Speech coding
  - Speaker recognition
- Summary and future directions

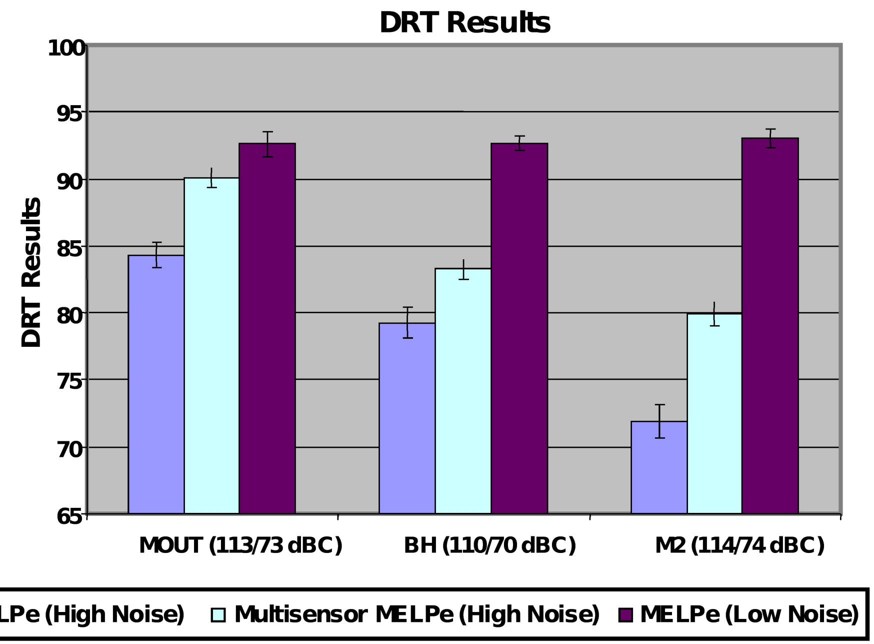
MELPe Speech Encoding
Multi-Sensor Fusion

Approach: MELPe signal enhancement with P-mic/Bone-mic/Res-mic signal fusion; GEMS pitch from MELPe analysis



DRT Intelligibility Test

Using 3 Males/3 Females from ASE corpus

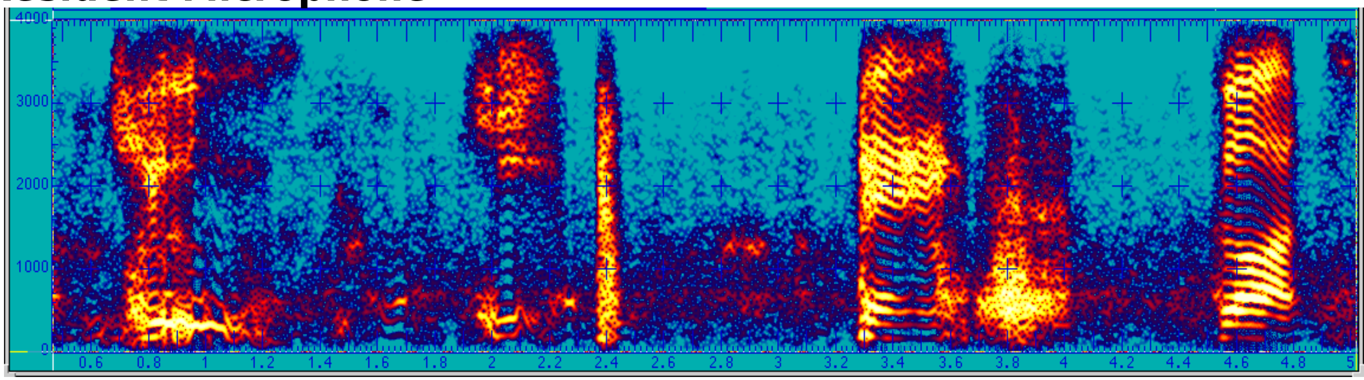


Significant intelligibility gains have been achieved in all of the high noise environments by exploiting ASE sensors (GEMS, P-mic, and Bone-mic).

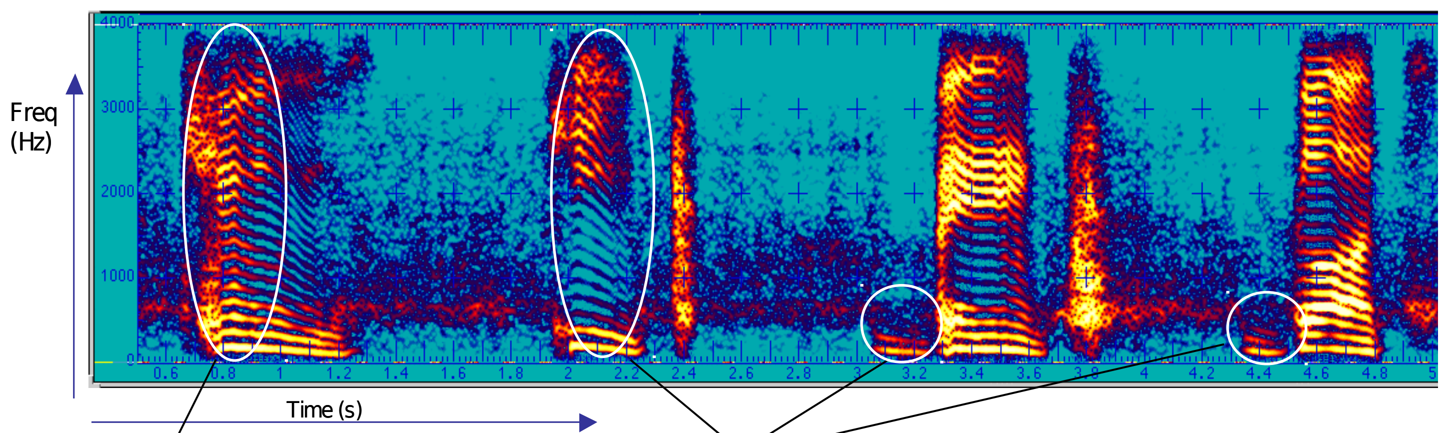
Multisensor MELPe
Demonstration

400 bps MELP coded speech in Bradley high-noise environment

Resident Microphone



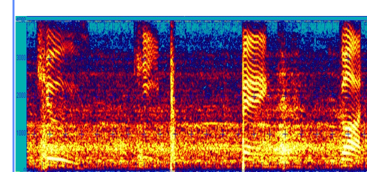
Multi-Sensor Enhancement



GEMS pitch helps to improve harmonic structure

P-mic low-band provides voice bars in voiced plosives and improved pitch in voicing

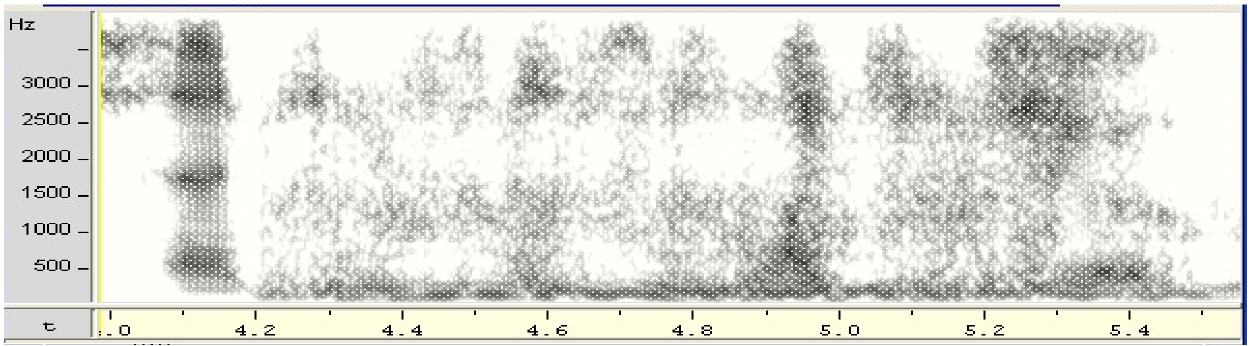
Original Noisy



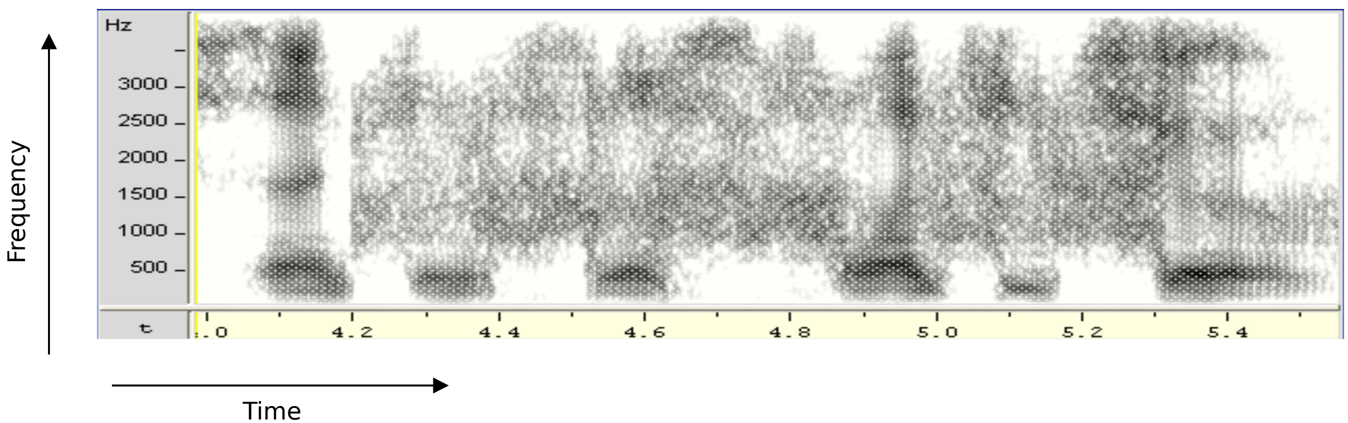
Multisensor MELPe
Demonstration

00 bps MELP coded speech in Military Urban high-noise environment

Resident Microphone



Multi-Sensor Enhancement



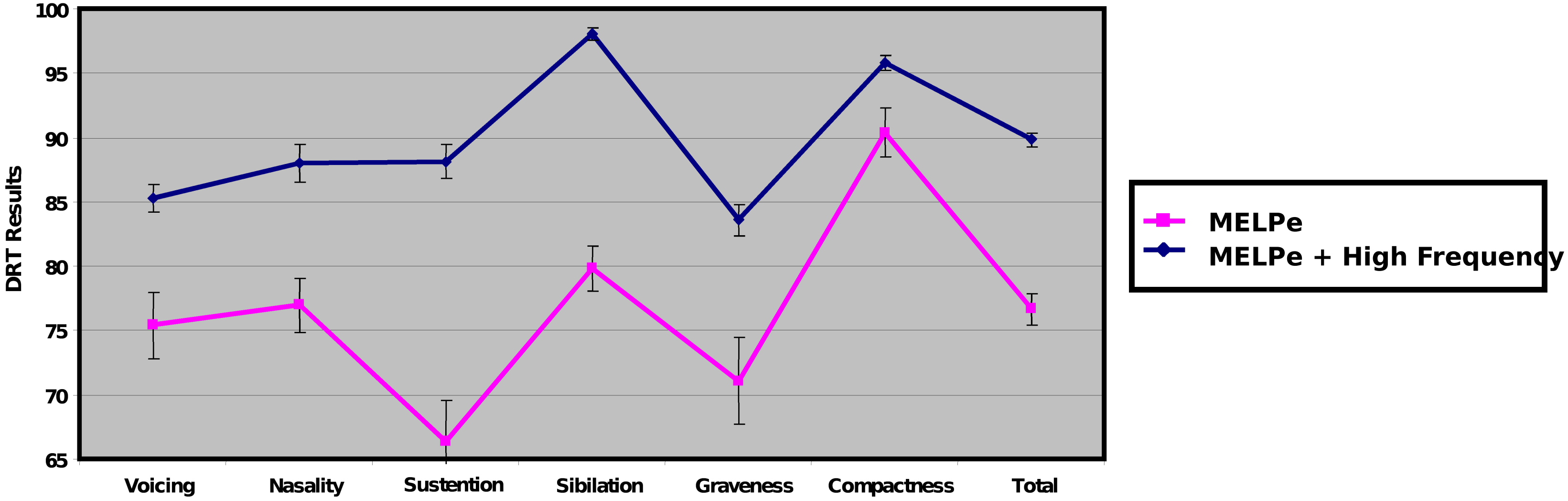
MELPe Speech Encoding
DRT Attribute Results

There is broad variation in the impact of ASE technology on various D intelligibility attributes --- with strong improvements in voicing and n

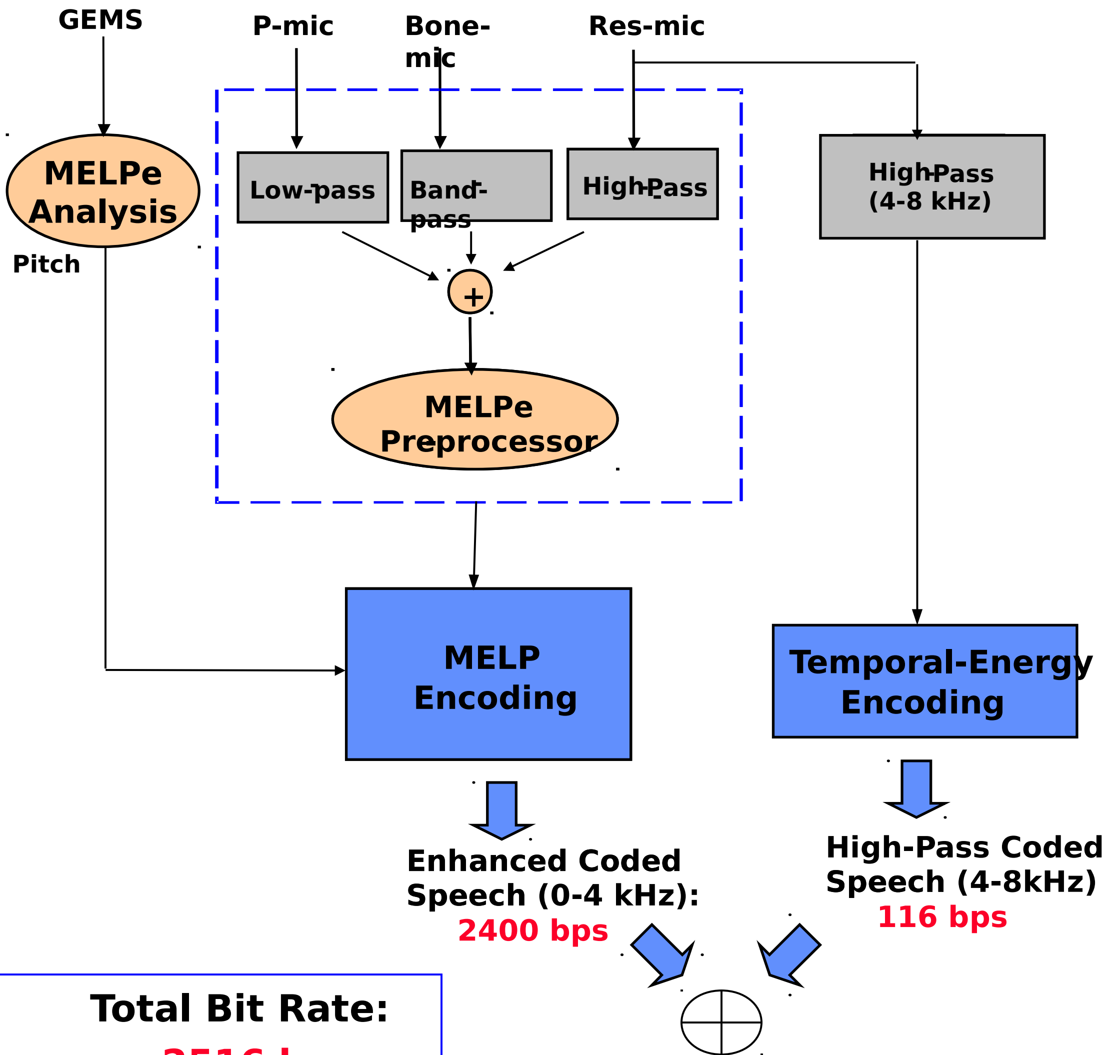
High Frequency Fusion
M2 High Noise Condition

- **High frequency (>4 kHz) speech data has been shown to provide significant intelligibility content**
  - **MELPe coded speech was augmented with high frequency unencoded speech**
    - **High frequency unencoded speech (4-8 kHz) was attenuated 100 dB in 0-4 kHz band**

- **Note that ARCON sound simulation rolls off at 4 kHz**
  - **Bradley Vehicle and Military Urban can exceed this range**



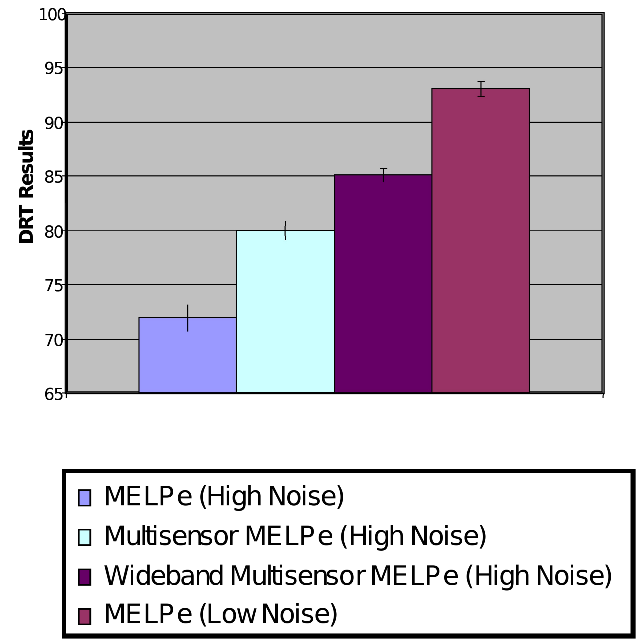
Wideband Multisensor MELPe



Total Bit Rate:
2516 bps

DRT Intelligibility Test

M2 Noise Environment



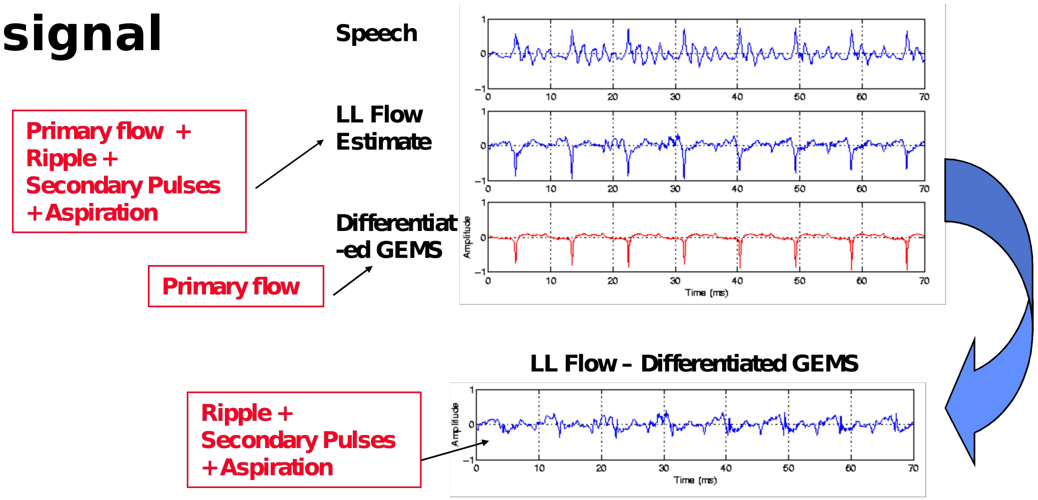
The addition of high frequency content to the Multisensor MELPe architecture provides significant DRT intelligibility gain.

Nonacoustic Sensors in Speaker Recognition

Motivation

- **Lincoln glottal flow estimator developed in late 90's**
  - **Pitch-synchronous inverse-filtering approach**
  - **Significant speaker ID in flow but not robust**

- **Comparison of Lincoln pitch-synchronous glottal flow estimator with differentiated GEMSsignal**



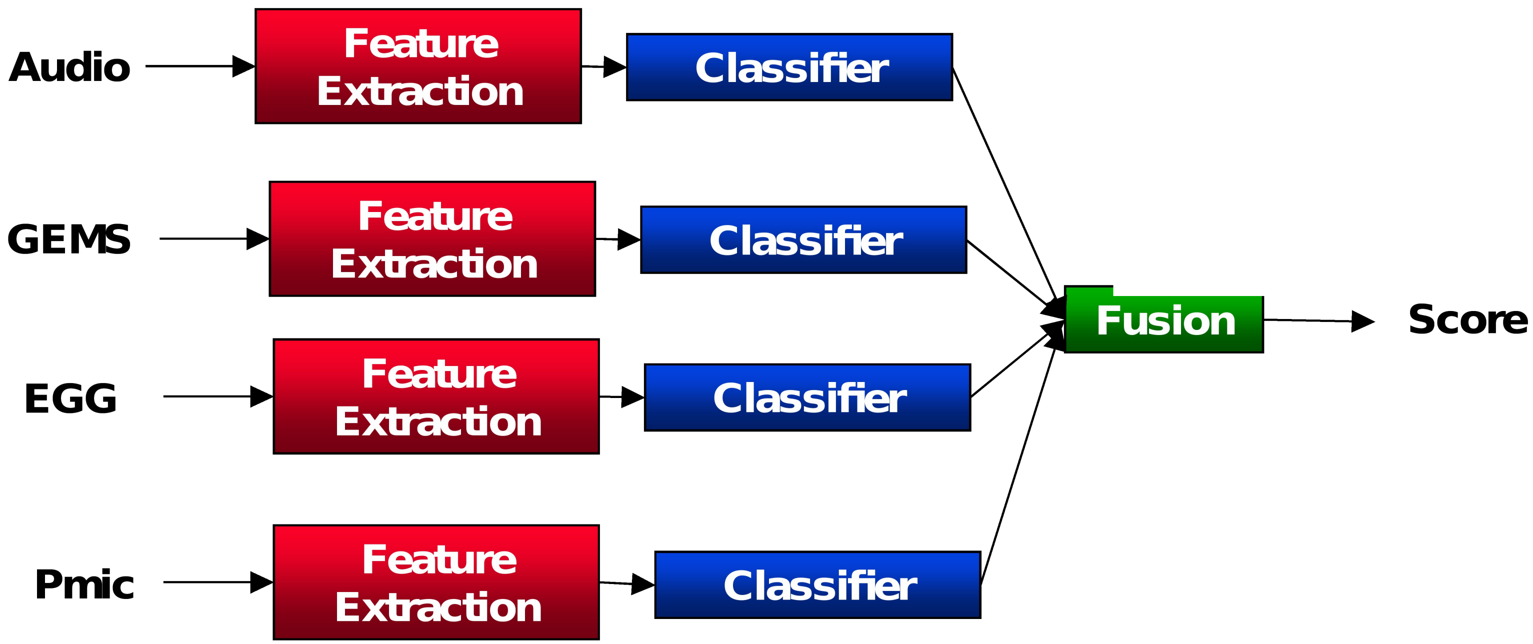
Needed: More exhaustive and quantitative study of relation of GEMS to acoustic-based flow estimate; Determination of "truth".

Speaker Recognition
Approach

Approach

- Treat each sensor output as we do a speech signal
- Apply standard feature extraction and classification
- Fuse at the score level

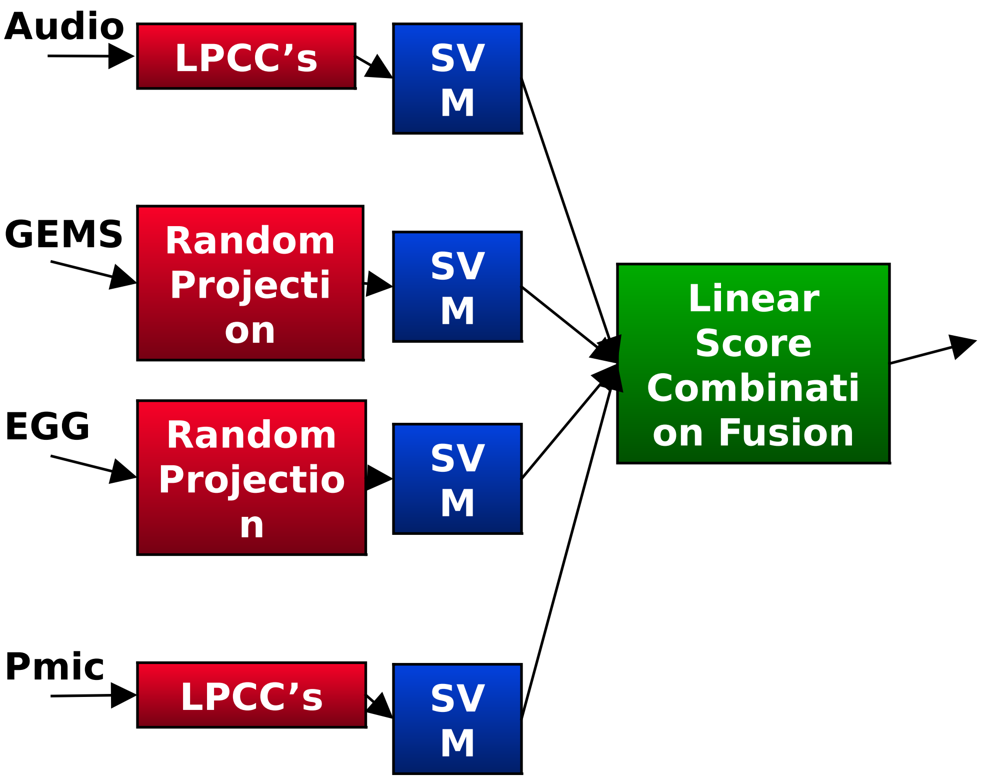
Multisensor Architecture



Speaker Recognition

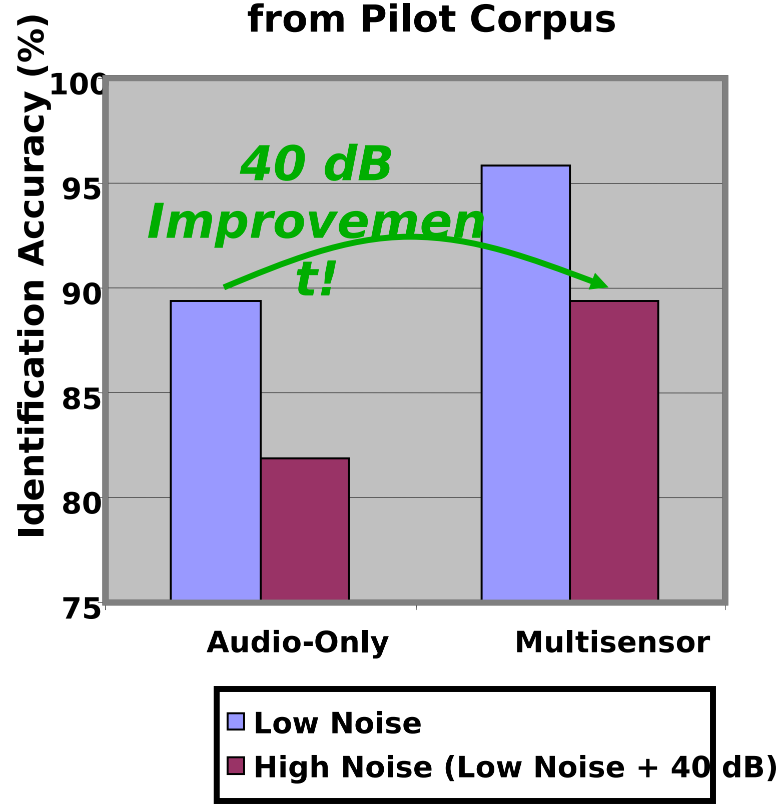
Results

Specific Multisensor Architecture



Speaker ID Results

**Limited training and testin
from Pilot Corpus**



**Speaker ID results using multiple sensors in a high noise environment mat
performance level of an audio-only approach in a low noise environment.**

Conclusions

- **Nonacoustic sensor measurements, such as from the GEMS and P-mic, have interesting properties**
  - **Reveal certain speech events such as voice bars and glottalized activity lost in the acoustic signal**
  - **Signal quality of different sensor outputs is band-dependent**

- **Nonacoustic sensors can be used with acoustic noise canceling microphones, such as resident microphones, to improve speech encoding and speaker recognition**
  - **Encoding: Primary gains in DRT attributes of voicing and nasality attributes**
    - **Corresponds to low-frequency and harmonic content of GEMS, P-mic, and bone-conduction-mic sensors**
  - **Speaker recognition: Large gain from fusion using standard recognition**

Some Key References

Lincoln Work

- D. Messing, Noise suppression using spectral magnitude and phase from nonacoustic sensors, MS Thesis, MIT, August 2003.
- T. F. Quatieri, D. Messing, K. Brady, W. B. Campbell, J. P. Campbell, M. Brandstein, C. J. Weinstein, J. D. Tardelli, and P. D. Gatewood, "Exploiting nonacoustic sensors for speech enhancement", Proc. Workshop on Multimodal User Authentication, Santa Barbara, CA, 11-12 December 2003.
- W.M. Campbell, T.F. Quatieri, J.P. Campbell, and C.J. Weinstein, "Multimodal speaker authentication using nonacoustic sensors," Proc. Workshop on Multimodal User Authentication, Santa Barbara, CA, 2003.
- K. Brady, T. F. Quatieri, W. B. Campbell, J. P. Campbell, M. Brandstein, C. J. Weinstein, "Multisensor MELPe using parameter substitution", Proc. Int. Conf. Acoustics, Speech, and Signal Processing, Montreal Canada, 2004.

Nonacoustic Sensors

- G.C. Burnett, J.F. Holzrichter, T.J. Gable, and L.C. Ng, "The use of glottal electromagnetic micropower sensors (GEMS) in determining a voiced excitation function," presented at the 138th Meeting of the Acoustical Society of America, November 2, 1999, Columbus, Ohio.
- M. Rothenberg, "A multichannel electroglottograph," J. of Voice, vol. 6, no. 1, pp. 36-43, 1992.
- M.V. Scanlon, "Acoustic sensor for health status monitoring," Proceedings of IRIS Acoustic and Seismic Sensing, vol. 2, pp. 205-222, 1998.
- T. Yanagisawa and K. Furihata, "Pickup of speech signal utilization of vibration transducer under high ambient noise", Journal of Acoustical Society of Japan, Vol. 31, No. 3, pp. 213-220, 1975.

2400 bps MELP

- A. McCree, K. Truong, E.B. George, T.P. Barnwell, and V, Viswanathan, "A 2.4 kbit/s MELP coder candidate for the new US Federal standard," Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Atlanta, GA, vol. 1, pp. 200-203, May 1996.